Search Snippet Evaluation at Yandex: Lessons Learned and Future Directions

Denis Savenkov, Pavel Braslavski, Mikhail Lebedev

Yandex
16, Leo Tolstoy St., Moscow 119021
{denxx, pb, mlebedev}@yandex-team.ru

Abstract. This papers surveys different approaches to evaluation of web search summaries and describes experiments conducted at Yandex. We hypothesize that the complex task of snippet evaluation is best solved with a range of different methods. Automation of evaluation based on available manual assessments and clickthrough analysis is a promising direction.

Keywords: evaluation, snippets, search summaries, web search, experimentation.

1 Introduction

A list of ranked document summaries is de facto a standard for web search result representation. A search summary[1] commonly consists of document title, original document fragments (namely text *snippets*), and metadata such as document date, size, URL, etc. Now we can observe the tendency of enriching web search summaries with images, so called QuickLinks, links to maps (e.g. in case the retrieved document is a company or organization homepage), user ratings of different kind, and other clues. Most text snippets originate from the original document and contain highlighted terms from the initial user query or their derivatives. Some snippets are, in fact, manually-crafted summaries from third-party sites (such as ODP[2] descriptions) or from META field of the original HTML page. A wide use of *microformats*[3] shifts the emphasis from the methods of choosing the best fragments from the original text to deciding whether to use the semantic mark-up provided by the page owner or not.

In some cases summaries can provide the user with the required information in situ (e.g. factoid questions). However, the main purpose of a search summary is to inform the user about the degree of relevance of the original retrieved document. Many studies confirm that search summaries have a big impact on the perceived search

[1] Also referred to as *result summary*, *snippet*, *query-biased summary*, *caption*, and *document surrogate*.
[2] http://dmoz.org/
[3] http://microformats.org/

quality of search: the user is unlikely to click on a misleading summary of a relevant document and, conversely, the user will be disappointed by a non-relevant document, if the summary suggested the opposite (however, the latter is a much less critical case). Turpin et al. [18] investigated how accounting for summary judgment stage can alter IR systems evaluation and comparison results. Based on a small user study, authors estimate that "14% of highly relevant and 31% of relevant documents are never examined because their summary is judged irrelevant" [18].

Web summary evaluation differs from search quality evaluation for several reasons. First, the notion of a "good summary" is multifaceted and sometimes contradictory. It is often hard to balance out different requirements. E.g. a snippet containing many query terms from different fragments of the original document is, in general, less readable. Longer snippets bear more information about the retrieved document but hinder overall comprehension of the search engine results page (SERP), etc. Second, summary judgments are only partially reusable (changes in generation algorithm lead to changes in an arbitrary subset of snippets for given query-document pairs).

In the industrial settings snippet evaluation can be aimed at different goals: 1) comparison with competitors, 2) evaluation of a new versions of snippet generation algorithm against production version, and 3) evaluation in favor of machine-learned algorithms for snippet generation.

In the next section we survey different approaches to search summaries evaluation and work in related areas. Section 3 describes different techniques used for snippet evaluation at Yandex, a Russian web search engine serving about 120M queries daily: an exploratory eye-tracking experiment, manual assessment of search snippets in terms of informativity and readability, automatic metrics, and evaluation based on clickthrough mining. Section 4 concludes and outlines the directions for further research.

2 Related Work

Snippet generation can be seen as a variant of general summarization task. There are two main approaches to summarization evaluation: 1) comparison against a gold standard or 2) task-oriented evaluation. Additionally, some intrinsic aspects of summaries such as readability or grammaticality are evaluated. Concurrent comparison, or side-by-side evaluation, of several summary variants is another option.

There are some approaches implemented within a series of standalone experiments or within evaluation campaigns of a larger scale.

In their pioneering work Tombros & Sanderson [17] compared the utility of query-biased summaries against first few sentences of retrieved documents in search results. A user study with 20 participants was performed on TREC *ad hoc* track data, i.e. topics and judged documents. Precision and recall of relevance judgments on summaries vs. leading sentences compared to available full document judgments, speed of judgments, and the need to refer to the full text were the indicators of the search results representation quality.

The task-oriented approach by White et al. [20] is in principal similar to one by Tombros & Sanderson. However, they tried to make search tasks closer to a real-world scenario and obtain a richer feedback from the users. 24 participants in the user study were asked to complete different search tasks using four different web search systems. Researchers used detailed questionnaires, accompanied by think-aloud, informal discussions, and automatic logging of users' actions during the experiment. The questionnaires contained the following statements regarding summary quality to be rated by participants: *The abstracts/summaries helped me to assess the pages for relevance; The abstracts/summaries showed my query terms in context*. The main automatic measure was the time spent on tasks.

Eye-tracking is a promising technique for testing user interfaces, including search results representation. Eye-tracking was used for investigation how snippet length affected user performance on navigational and informational search tasks [4]. The main findings are that longer snippets improved performance for informational queries but worsened it for navigational queries. Eye-tracking allowed to support these conclusions, i.e. a longer snippet distracted the user's attention from the URL line. The study [10] supports findings that different query types are best answered by snippets of different length. Leal Bando et al. [12] used eye-tracking in a small user study (four query-document pairs, 10 participants) to juxtapose document's fragments used by humans for generative vs. extractive query-biased summaries und showed that humans focused on the same pieces of text for both tasks most of the time. Comparison of automatically generated against human-crafted snippets suggested that gold-standard evaluation must account not only for word overlap but also for position information.

Mechanical Turk[4] crowdsourcing was used in a study on temporal snippets [2]. Mechanical Turk judges, presented with three variants of snippets for a Wikipedia page at once, had to choose the best one and provide additional response. 30 snippets corresponding to 10 queries were evaluated in total.

Clarke et al. studied snippet features that potentially influenced snippet quality and consequently – user behavior [3]. The authors performed clickthrough mining of a commercial search engine. In contrast to previous work based on rather small user studies, this study enabled a large-scale experiment in a less artificial setting. The authors looked at *clickthrough inversions* as a signal of snippet attractiveness: the pairs of consequent snippets in result list, where the lower result received more clicks than the higher-ranked one. The study confirmed the perception that the presence of query terms in a snippet, its length, complexity of URL, and readability contribute to overall quality of snippets.

Kanungo & Orr [11] reported on a machine-learned readability measure for search snippets. The model was trained on about 5,000 human judgments and incorporated 13 various features such as *average characters per word, percentage of complex words, number of fragments, query word hit fraction* etc. The trained model predicted human judgment well and can be used both for continuous large-scale monitoring of snippet readability and for improving existing summarizers.

---

[4] http://www.mturk.com/

DUC/TAC series of workshops[5] has been focusing on evaluation methodology for automatic summarization for several years. The initiative collected a sizeable volume of system-produced summaries, ideal human-crafted summaries, and comparisons of system summaries with ideal summaries performed by humans. These data enabled the introduction of automatic quality measures based on proximity of an automatically generated summary and a set of ideal summaries. Proximity can be defined in terms of common n-grams, word sequences, or similar syntactic units. ROUGE [13] and Basic Elements (BE) [6] exemplify these approaches and show a good correlation with systems rankings based on human judgments. Automatic measure allows re-using of judgments.

The last edition of the TAC multidocument summarization included 46 topics for guided summarization. The task was to produce a 100-word summary from the first 10 documents on a certain topic and an update summary for the second 10 documents. Automatically generated summaries were evaluated and compared to ideal summaries by human judges in respect of responsiveness (relevance to topic), readability, and Pyramid (content similarity to human summaries) [14]. In contrast to web queries, the task presents a much more detailed description of the information need, its aspects, and prior knowledge on the topic.

Snippet generation can be seen as passage retrieval task, i.e. retrieving the fragments of a document relevant to a particular information need. Passage retrieval task was evaluated within TREC HARD track in 2003[5] and 2004. System results were evaluated against fragments of documents marked as relevant by annotators. How to quantify the character-level overlap of ideal fragments with systems' output is discussed in [19].

Two years (2007 & 2008) WebCLEF[6] offered snippet generation/information synthesis task: participants were presented with a topic description and up to 100 Google results to relevant search queries. A system response was a ranked list of plain text snippets extracted from the retrieved documents (first 7,000 characters of the system response were assessed). System responses were pooled, and assessors were asked to mark text spans with useful information. Average character precision and average character recall were used for evaluation similarly to TREC HARD track. ROUGE-1 and ROUGE-1-2 turned out to be not quite appropriate for evaluation of the task [9, 15].

Recently INEX announced a snippet evaluation track [7]. The task is to return snippets limited to 300 characters for retrieved Wikipedia articles. Evaluation metrics will employ comparison of relevance assessments based on whole documents vs. short snippets.

1CLICK subtask of the NTCIR-9 Intent task [1] is running at the time of writing (June 2011). It resembles snippet generation, QA, and information synthesis tasks: for a given query the system must return a string of 140 ('mobile run') or 500 ('desktop run') characters. A Japanese collection and Japanese queries are used. Evaluation is

[5] http://duc.nist.gov/, http://www.nist.gov/tac/

[6] http://ilps.science.uva.nl/WebCLEF/

based on information nuggets presented in the system's response (similar to content similarity in TAC evaluation).

3 Snippet Evaluation at Yandex

In order to establish a snippet evaluation routine at Yandex, we experimented with a wide range of techniques and approaches in line with those described in Section 2: pairwise comparison of two versions/systems, relevance on whole documents vs. snippets, direct readability assessment, clickthrough mining, etc. The work is still in progress. Our current perception is that it is very hard to invent an integral measure of snippet quality. Thus, we suggest using a set of different tools and approaches for different aspects and goals of snippet evaluation.

3.1 Eye Tracking Experiment

Eye-tracking became very popular for investigating user behavior and usability of user interfaces. We employed eye-tracking for better understanding of how different aspects of snippet quality influence user satisfaction. One of the research questions was whether highlighting additional terms reflecting possible user intents was helpful.

We prepared 19 tasks of different types, e.g. download a given popular song, find information for writing an essay on a given topic, find the address of a given movie theatre, find term definition, etc. Some tasks were open, while for others initial search queries were provided. 20 participants took part in the study, each participant was allotted an hour to complete the tasks. Both experienced and beginner, frequent and occasional Yandex users took part in the study. Participants were divided into two groups – the first group was presented with standard snippets, the second group had snippets with terms related to the query intent (e.g. "buy" for commercial queries) highlighted along with the query terms.

The main conclusions from our user study are as follows:

1. The title is much more important than the body of the snippet. Users skip relevant results with no highlighted terms in the title in favor of lower-ranked results with seemingly better titles.
2. Highlighting attracts users' attention and helps them navigate through the results list. Users click directly on highlighted terms in the snippet titles. Additional highlighted terms, e.g. reflecting query intents, help users find the answer faster and draw their attention to results in the lower part of SERP (supports [8], see Fig.1).
3. Experienced users prefer skimming: they examine snippet fragments around highlighted words, jumping from one part of the snippet to another. If the title contains relevant information, these users prefer clicking on the link without examining the body of the snippet.
4. Users rely on ranking – high-ranked results are clicked regardless of the snippet's content or quality (supported by many click-log experiments). However, some

users get bored after examining a few results at the top of the result list and scroll down to the lower part.

5. Inexperienced users are somewhat "scared to click"; they usually examine a considerable number of results before clicking. Novices examine snippet content more thoroughly before moving on to the next result.
6. Users go to the original document, even if the snippet contains a complete answer to their factoid query (supports [2]).
7. Some users are conservative and shy away from a certain type of snippets, e.g. containing image or video thumbnails.

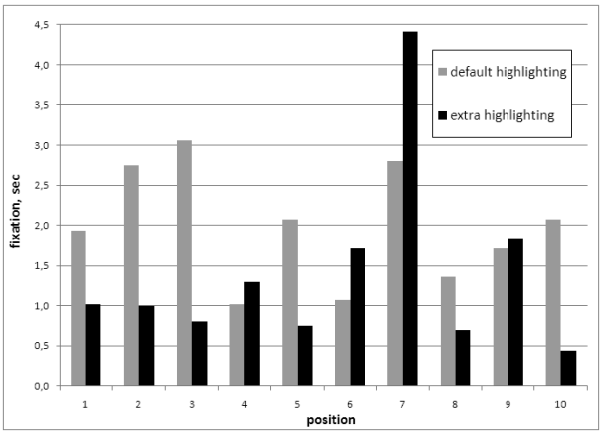

Fig. 1. Averaged fixation times for two groups (10 participants each) solving the same task: 1) default highlighting – query terms only; 2) query intents ("download", in this case) additionally set off in bold in snippets at positions 7 and 9

3.2 Manual Assessment

Manual assessment performed by trained judges is the basis of traditional evaluation methodology in the field of information retrieval. We performed a sizeable manual assessment within the experiments on machine-learned snippets.

The key features of an ideal snippet are: 1) it conveys sufficient information about the whole document in the context of a query (i.e. users can assess the document's relevance to the query based on a snippet); 2) it is easy to read/understand. These qualities are reflected in the snippet *informativity* and *readability* measures.

During initial experiments we realized that it was hard for an assessor to score a snippets' informativity on an absolute scale. Even if we have absolute scores it is questionable whether these scores are comparable across different queries. A much easier task is to compare and rank different snippet variants for a given query-document pair.

The interface of the assessment tool is presented in Fig. 2. Assessor is presented with a query and a randomly ordered list of up to 10 snippet variants for the same document produced by different snippet generation algorithms. Query-document pairs

were sampled so that their relevance distribution was close to Yandex's results. The task was to move individual snippets up and down and insert "borderlines", thus, creating ranked groups of snippets of an approximately equal quality. Evaluation for informativity and readability was performed independently (different assessors ranked the same task for informativity and for readability). The study was performed in three stages; 11 judges participated in the study. Table 1 describes some statistics of the evaluation process. It is interesting to observe a learning effect in informativity evaluation in terms of speed and quality: average time spent on task decreases, as well as the proportion of tasks with at least two candidates reverse-ordered compared to tasks evaluated by assessors' supervisor. Time spent on readability evaluation does not show this behavior and rather correlates with the average snippet length.

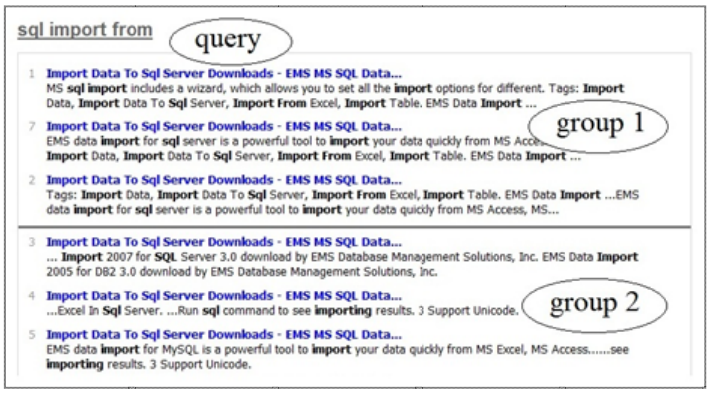

Fig. 2. Tool for relative assessment of different snippet variants for a query-document pair

Table 1. Statistics of manual evaluation experiment

Period	Query-doc pairs	Ave. snippet length	Time spent on inform. task, sec	Reverse-ordered pairs, inform., %	Time spent on readab. task, sec
Mar 2010	1,200	250	115	29	72
Jun 2010	3,200	170	107	26	60
Jan 2011	2,000	250	101	24	84

Based on a subset of evaluation results, we calculated Kendall tau-b correlation between informativity and readability rankings (Table 2). One can see that the correlation is positive, i.e. snippets tend to be good or bad in both aspects. However, the correlation is low, which implies that we have to consider readability and informativity as complementary snippet features. It is interesting to notice that informativity and readability are less correlated in long queries. For short queries a readable text fragment containing one or two query terms is more likely to contain useful information; for longer queries this dependency becomes more complex.

It is worth to mention that the same evaluation guidelines (but different interface) were used by assessors performing "blind" side-by-side comparison of snippets on Yandex against competitors.

Table 2. Correlation between readability and informativity

Query length	# of query-doc pairs	# of snippets	*r*
1	164	1,481	0.432
2	256	2,266	0.401
3	273	2,466	0.374
4	183	1,588	0.363
≥5	237	2,024	0.353
total	1,113	9,825	0.383

3.3 Automated Quality Measures

Manual assessment is very expensive and time-consuming even considering the availability of services like Mechanical Turk. When changing the snippet generation algorithm, we need a simple and fast method to assess the new version. At the moment, we use a range of automated measures that capture some snippet features:

- General number of highlighted terms, proportion of query terms presented in the snippet, proportion of highlighted terms and their variations, such as density and diversity of highlighted terms, number of highlighted terms in title, etc.
- Snippet's 'neatness', which is closely related to its readability. We measure the number of non-readable characters (#, %, ^, @, *, <, >, etc.), the number of porn words, etc.
- The number of 'empty' snippets (i.e. title-only snippets).

Table 3 presents Kendall tau-b correlation between some automated measures and manual rankings of snippets regarding informativity and readability (calculated on the same data as in Table 2).

Table 3. Correlation between assessors' rankings and rankings based on automated measures

Query length	Informativity vs.		Readability vs.	
	proportion of query terms	# of highlighted terms	proportion of non-readable chars	# of fragments
1	0.205	0.206	–0.322	–0.699
2	0.281	0.329	–0.304	–0.695
3	0.302	0.403	–0.309	–0.671
4	0.328	0.484	–0.327	–0.641
≥5	0.334	0.535	–0.323	–0.576
Total	0.274	0.424	–0.306	–0.657

As expected, the proportion of query terms presented in a snippet and the number of highlighted terms positively correlated with informativity, whereas the proportion of non-readable characters and the number of fragments from the original document in a snippet negatively correlated with readability. However, the correlation is not strong, except for the number of fragments.

In addition, Table 4 presents some automated measures for two snippet generation algorithms produced during routine development at Yandex. In general, *Alg2* shows a better behavior, the only drawback is a slightly increased number of non-readable characters.

Table 4. Automated measures for two snippet generation algorithms (2,000 queries, 17,009 snippets generated by each algorithm)

Measure	Alg1	Alg2
Proportion of query terms in snippets	0.762	0.774
Proportion of snippets containing all query terms	0.550	0.568
Snippet length in chars	165.76	161.59
# of highlighted query terms per snippet	3.317	3.368
Proportion of non-readable chars	0.020	0.022
Average word length	5.901	5.870

3.4 A/B Testing

Automatic evaluation of information retrieval systems based on user behavior is an area of active research. Automatic methods promise to make evaluation faster, cheaper, and more representative. However, despite that a plethora of data is available, the crucial problem remains interpreting these data in terms of quality.

We perform automatic evaluation of a new candidate snippet generation algorithm against the production version using A/B testing. A subset of user population is presented with search results with the same ranking but featuring different snippets. In general, we used a subset of metrics described in [16] (session-based metrics, such as *queries per session* or *reformulation rate* are not quite appropriate for snippet evaluation). However, in contrast to ranking evaluation, some metrics receive a different interpretation. For example, an increased CTR of lower positions in case of shorter snippets can indicate a positive change: the user develops a better general comprehension of SERP, whereas in case of ranking evaluation it might mean that good results are lower.

The main purpose of snippets is to help users find relevant documents on the search engine results page and avoid those that are irrelevant. Thus, the first important behavior characteristic is dwell time, i.e. the time the user spends on the web page after clicking the link on the search results page. The higher the proportion of SERP clicks with long dwell times is, the fewer documents with non-representative summaries there are in search results. Also, the less the abandonment rate (i.e. queries with no clicks on results) is, the better annotations the documents on SERP have. In addition, an increase of CTRs for the lower-ranked documents usually suggests that

the snippets for top-ranked documents get less attention because they are not informative enough (cf. click inversions [3]). But this depends highly on the length of snippets, since the shorter the snippets are, the higher CTRs the lower documents have. In addition to dwell time, we need to take into account the time required to find the answer to the user's query. For example, the time it takes to make the first click is a very useful measure, which correlates with the time it takes to find the answer.

Table 5. A/B testing results for two snippet generation algorithms (*statistically significant at the 0.01 confidence level)

Measure	Alg1	Alg2
Abandoned queries, %	38.270	38.220 (–0.13%)*
Click inversions, %	6.8017	6.8212 (+0.29%)*
Long dwell times rate, %	72.5897	72.6088 (+0.026%)
Time to first click, sec	11.5274	11.5245 (–0.02%)
1st position CTR	0.3786	0.3790 (+0.10%)*
2nd position CTR	0.1631	0.1630 (–0.03%)
9th position CTR	0.0355	0.0357 (+0.42%)*
10th position CTR	0.0358	0.0360 (+0.27%)*

Table 6. A/B testing results for snippets with extra highlighting of possible user intents (*statistically significant at the 0.01 confidence level)

Measure	Default highlighting	Extra highlighting
Abandoned queries, %	40.0031	39.9052 (–0.25%)*
Click inversions, %	6.4506	6.4818 (+0.48%)*
Long dwell times rate, %	73.8379	73.7960 (–0.06%)
Time to first click, sec	11.6832	11.6638 (–0.17%)*
1st position CTR	0.3132	0.3138 (+0.19%)*
2th position CTR	0.1639	0.1645 (+0.33%)*
9th position CTR	0.0343	0.0347 (+1.11%)*
10th position CTR	0.0422	0.0424 (+0.45%)*

Table 5 presents user behavior metrics for two different snippet generation algorithms (the same as in the previous section). *Alg2* snippets were shown to 12.5% of users during two weeks (May 10–24, 2011). Since snippets generated by *Alg2* contained more query terms and were slightly shorter, we could observe increased CTRs, especially for lower positions. Due to this fact, click inversion rate increased (more attention to lower positions resulted in more click inversions). More highlighting resulted in a lower number of abandoned queries. Proportion of long (>30 sec) dwell times for *Alg2* was approximately the same as for *Alg1*. This might mean that *Alg2* generated more attractive snippets for both relevant and non-relevant documents. Since the total number of clicks on the links to relevant documents increased, we could conclude that *Alg2* generated better snippets than *Alg1*.

Table 6 shows the results of another experiment for snippet generation algorithms that differ only in the way they highlighted terms. The experiment was performed on

50% of users for two weeks. Clickthrough mining supported the results of the eye-tracking experiment; it showed that increased attractiveness of snippets resulted in higher CTRs and shorter times to first click.

4 Conclusions and Future Research

Based on our experiments we can conclude that the complex and diverse task of snippet evaluation is best solved with a range of different methods – user studies, automated measures, manual evaluation, and clicktrough mining.

Thus, we use eye-tracking when introducing changes in general SERP layout or snippet representation: snippet length, snippets enriched by video and image thumbnails, QuickLinks, and links to maps, customized snippets for recipes, hotels, forums, and products, extra highlighting, URL representation, etc.

Manual evaluation is employed for machine-learned snippet generation and comparison with competitors. We use relative quality assessments for two aspects of snippets – informativity and readability. The main drawback of the approach is that judgments cannot be re-used. However, approaches that allowed for re-using of data – e.g. ideal snippets extracted by humans – are much more costly and time-consuming and presumably show less inter-annotator agreement.

Automatic measures are suitable for fast, albeit rough, evaluation of snippet generation algorithms. We use them as regression tests for newly developed algorithms. Moreover, we plan to implement automated measures based on manual readability evaluation results (in a way similar to [11]).

A/B testing is the final step in shipping snippet generation algorithm to production.

We now plan to address the problem of building an integral snippet evaluation metrics and automation of snippet metrics based on available manual assessment results and click data analysis. In addition, we plan to conduct a manual assessment of information nuggets presented in snippets for factoid queries (analogously to DUC/TAC/1CLICK approach).

5 References

1. 1CLICK@NTCIR-9, http://research.microsoft.com/en-us/people/tesakai/1click.aspx
2. Alonso, O., Baeza-Yates, R., Gertz, M.: Effectiveness of Temporal Snippets. In: WSSP Workshop at the World Wide Web Conference—WWW'09 (2009)
3. Clarke, Ch., Agichtein, E., S. Dumais, White, R. W.: The Influence of Caption Features on Clickthrough Patterns in Web Search. In: SIGIR2007 (2007)
4. Cutrell, E., Guan, Zh.; What Are You Looking For? An Eye-tracking Study of Information Usage in Web Search. In: CHI'07 (2007)
5. HARD, High Accuracy Retrieval from Documents. TREC 2003 track guidelines, http://ciir.cs.umass.edu/research/hard/guidelines2003.html
6. Hovy, E., Lin, Ch.-Y., Zhou, L.: Evaluating DUC 2005 Using Basic Elements. In: Fifth Document Understanding Conference (DUC), Vancouver, Canada (2005)
7. INEX 2011 Snippet Retrieval Track, https://inex.mmci.uni-saarland.de/tracks/snippet/

8. Iofciu, T., Craswell, N., Shokouhi, M.: Evaluating the Impact of Snippet Highlighting in Search. In: Understanding the User Workshop – SIGIR'09 (2009)
9. Jijkoun, V., de Rijke, M.: Overview of WebCLEF 2008. In: Evaluating Systems for Multilingual and Multimodal Information Access. LNCS, vol. 5706, pp. 787--793 (2009)
10. Kaisser, M., Hearst, M. A., Lowe, J. B.: Improving Search Results Quality by Customizing Summary Lengths. In: ACL-08: HLT (2008)
11. Kanungo, T., Orr, D.: Predicting the Readability of Short Web Summaries. In WSDM '09 (2009)
12. Leal Bando, L., Scholer, F., Turpin, A.: Constructing Query-biased Summaries: a Comparison of Human and System Generated Snippets. In: IIiX'2010 (2010)
13. Lin, Ch.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: ACL'04 Workshop: Text Summarization Branches Out, Barcelona, Spain (2004)
14. Nenkova, A., Passonneau, R. J., McKeown, K.: The Pyramid Method: Incorporating Human Content Selection Variation in Summarization Evaluation. In: TSLP 4(2) (2007)
15. Overwijk, A., Nguyen, D., Hauff, C., Trieschnigg, R., Hiemstra, D., de Jong, F.: On the Evaluation of Snippet Selection for WebCLEF. In: Evaluating Systems for Multilingual and Multimodal Information Access. LNCS, vol. 5706, pp. 794--797 (2009)
16. Radlinski, F., Kurup, M., Joachims, T.: How Does Clickthrough Data Reflect Retrieval Quality? In: CIKM'08 (2008)
17. Tombros, A., Sanderson, M.: Advantages of Query Biased Summaries in Information Retrieval. In: SIGIR'98 (1998)
18. Turpin, A., Scholer, F., Jarvelin, K., Wu, M., Culpepper, J.S.: Including Summaries in System Evaluations. In: SIGIR'09 (2009)
19. Wade, C., Allan, J.: Passage Retrieval and Evaluation. Technical report, CIIR, University of Massachusetts, Amherst (2005)
20. White, R. W., Jose, J. M., Ruthven I.: A Task-Oriented Study on the Influencing Effects of Query-Biased Summarisation in Web Searching. Information Processing and Management, 39 (2003)